

INTELLIGENT NETWORK INTERFACE SYSTEM
AND METHOD FOR ACCELERATED PROTOCOL PROCESSING

Inventors

5 Laurence B. Boucher, Clive M. Philbrick, Daryl D. Starr, Stephen E. J. Blightman, Peter K. Craft, and David A. Higgen.

Cross-Reference to Related Applications

The present application claims the benefit under 35 USC § 120 of (is a
10 continuation of) U.S. Patent Application Serial No. 09/802,550, filed March 9, 2001,
which claims the benefit under 35 USC § 120 of (is a continuation-in-part of) U.S. Patent
Application Serial No. 09/067,544, filed April 27, 1998, which claims the benefit under
35 USC § 119 of U.S. Patent Application Serial No. 60/061,809, filed October 14, 1997.
U.S. Patent Application Serial No. 09/802,550, filed March 9, 2001, also claims the
15 benefit under 35 USC § 120 of (is a continuation-in-part of) U.S. Patent Application
Serial No. 09/141,713, filed August 28, 1998, which claims the benefit under 35 USC §
119 of U.S. Patent Application Serial No. 60/098,296, filed August 27, 1998. U.S. Patent
Application Serial No. 09/802,550, filed March 9, 2001, also claims the benefit under 35
USC § 120 of (is a continuation-in-part of) U.S. Patent Application Serial No.
20 09/384,792, filed August 27, 1999, (is a continuation-in-part of) U.S. Patent Application
Serial No. 09/416,925, filed October 13, 1999, (is a continuation-in-part of) U.S. Patent
Application Serial No. 09/439,603, filed November 12, 1999, (is a continuation-in-part
of) U.S. Patent Application Serial No. 09/464,283, filed December 15, 1999, U.S. Patent
Application Serial No. 09/514,425, filed February 28, 2000, (is a continuation-in-part of)
25 U.S. Patent Application Serial No. 09/675,484, filed September 29, 2000, (is a

continuation-in-part of) U.S. Patent Application Serial No. 09/675,700, filed September 29, 2000, (is a continuation-in-part of) U.S. Patent Application Serial No. 09/692,561, filed October 18, 2000, (is a continuation-in-part of) U.S. Patent Application Serial No. 09/748,936, filed December 26, 2000, (is a continuation-in-part of) U.S. Patent

5 Application Serial No. 09/789,366, filed February 20, 2001, (is a continuation-in-part of) U.S. Patent Application Serial No. 09/801,488, filed March 7, 2001, (is a continuation-in-part of) U.S. Patent Application Serial No. 09/802,551, filed March 9, 2001, (is a continuation-in-part of) U.S. Patent Application Serial No. 09/802,426, filed March 9, 2001, and (is a continuation-in-part of) U.S. Patent Application Serial No. 10/005,536, filed November 7, 2001. All of the applications listed above are incorporated by reference herein.

10

Technical Field

The present invention relates generally to computer or other networks, and more particularly to protocol processing for information communicated between hosts such as computers connected to a network.

15

Background

The advantages of network computing are increasingly evident. The convenience and efficiency of providing information, communication or computational power to individuals at their personal computer or other end user devices has led to rapid growth of such network computing, including internet as well as intranet systems and applications.

20

As is well known, most network computer communication is accomplished with the aid of a layered software architecture for moving information between host computers connected to the network. The layers help to segregate information into manageable segments, the general functions of each layer often based on an international standard

5 called Open Systems Interconnection (OSI). OSI sets forth seven processing layers through which information may pass when received by a host in order to be presentable to an end user. Similarly, transmission of information from a host to the network may pass through those seven processing layers in reverse order. Each step of processing and service by a layer may include copying the processed information. Another reference

10 model that is widely implemented, called TCP/IP (TCP stands for transport control protocol, while IP denotes internet protocol) essentially employs five of the seven layers of OSI.

Networks may include, for instance, a high-speed bus such as an Ethernet connection or an internet connection between disparate local area networks (LANs), each

15 of which includes multiple hosts, or any of a variety of other known means for data transfer between hosts. According to the OSI standard, physical layers are connected to the network at respective hosts, the physical layers providing transmission and receipt of raw data bits via the network. A data link layer is serviced by the physical layer of each host, the data link layers providing frame division and error correction to the data

20 received from the physical layers, as well as processing acknowledgment frames sent by the receiving host. A network layer of each host is serviced by respective data link layers, the network layers primarily controlling size and coordination of subnets of packets of data.

A transport layer is serviced by each network layer and a session layer is serviced by each transport layer within each host. Transport layers accept data from their respective session layers and split the data into smaller units for transmission to the other host's transport layer, which concatenates the data for presentation to respective presentation layers. Session layers allow for enhanced communication control between the hosts. Presentation layers are serviced by their respective session layers, the presentation layers translating between data semantics and syntax which may be peculiar to each host and standardized structures of data representation. Compression and/or encryption of data may also be accomplished at the presentation level. Application layers are serviced by respective presentation layers, the application layers translating between programs particular to individual hosts and standardized programs for presentation to either an application or an end user. The TCP/IP standard includes the lower four layers and application layers, but integrates the functions of session layers and presentation layers into adjacent layers. Generally speaking, application, presentation and session layers are defined as upper layers, while transport, network and data link layers are defined as lower layers.

The rules and conventions for each layer are called the protocol of that layer, and since the protocols and general functions of each layer are roughly equivalent in various hosts, it is useful to think of communication occurring directly between identical layers of different hosts, even though these peer layers do not directly communicate without information transferring sequentially through each layer below. Each lower layer performs a service for the layer immediately above it to help with processing the communicated information. Each layer saves the information for processing and service

to the next layer. Due to the multiplicity of hardware and software architectures, systems and programs commonly employed, each layer is necessary to insure that the data can make it to the intended destination in the appropriate form, regardless of variations in hardware and software that may intervene.

5 In preparing data for transmission from a first to a second host, some control data is added at each layer of the first host regarding the protocol of that layer, the control data being indistinguishable from the original (payload) data for all lower layers of that host. Thus an application layer attaches an application header to the payload data and sends the combined data to the presentation layer of the sending host, which receives the combined 10 data, operates on it and adds a presentation header to the data, resulting in another combined data packet. The data resulting from combination of payload data, application header and presentation header is then passed to the session layer, which performs required operations including attaching a session header to the data and presenting the resulting combination of data to the transport layer. This process continues as the 15 information moves to lower layers, with a transport header, network header and data link header and trailer attached to the data at each of those layers, with each step typically including data moving and copying, before sending the data as bit packets over the network to the second host.

 The receiving host generally performs the converse of the above-described 20 process, beginning with receiving the bits from the network, as headers are removed and data processed in order from the lowest (physical) layer to the highest (application) layer before transmission to a destination of the receiving host. Each layer of the receiving host recognizes and manipulates only the headers associated with that layer, since to that

layer the higher layer control data is included with and indistinguishable from the payload data. Multiple interrupts, valuable central processing unit (CPU) processing time and repeated data copies may also be necessary for the receiving host to place the data in an appropriate form at its intended destination.

5 The above description of layered protocol processing is simplified, as college-level textbooks devoted primarily to this subject are available, such as Computer Networks, Third Edition (1996) by Andrew S. Tanenbaum, which is incorporated herein by reference. As defined in that book, a computer network is an interconnected collection of autonomous computers, such as internet and intranet systems, including
10 local area networks (LANs), wide area networks (WANs), asynchronous transfer mode (ATM), ring or token ring, wired, wireless, satellite or other means for providing communication capability between separate processors. A computer is defined herein to include a device having both logic and memory functions for processing data, while computers or hosts connected to a network are said to be heterogeneous if they function
15 according to different operating systems or communicate via different architectures.

As networks grow increasingly popular and the information communicated thereby becomes increasingly complex and copious, the need for such protocol processing has increased. It is estimated that a large fraction of the processing power of a host CPU may be devoted to controlling protocol processes, diminishing the ability of
20 that CPU to perform other tasks. Network interface cards have been developed to help with the lowest layers, such as the physical and data link layers. It is also possible to increase protocol processing speed by simply adding more processing power or CPUs according to conventional arrangements. This solution, however, is both awkward and

expensive. But the complexities presented by various networks, protocols, architectures, operating systems and applications generally require extensive processing to afford communication capability between various network hosts.

5 Summary

The current invention provides a system for processing network communication that greatly increases the speed of that processing and the efficiency of moving the data being communicated. The invention has been achieved by questioning the long-standing practice of performing multilayered protocol processing on a general-purpose processor.

10 The protocol processing method and architecture that results effectively collapses the layers of a connection-based, layered architecture such as TCP/IP into a single wider layer which is able to send network data more directly to and from a desired location or buffer on a host. This accelerated processing is provided to a host for both transmitting and receiving data, and so improves performance whether one or both hosts involved in
15 an exchange of information have such a feature.

The accelerated processing includes employing representative control instructions for a given message that allow data from the message to be processed via a fast-path which accesses message data directly at its source or delivers it directly to its intended destination. This fast-path bypasses conventional protocol processing of headers that
20 accompany the data. The fast-path employs a specialized microprocessor designed for processing network communication, avoiding the delays and pitfalls of conventional software layer processing, such as repeated copying and interrupts to the CPU. In effect, the fast-path replaces the states that are traditionally found in several layers of a

conventional network stack with a single state machine encompassing all those layers, in contrast to conventional rules that require rigorous differentiation and separation of protocol layers. The host retains a sequential protocol processing stack which can be employed for setting up a fast-path connection or processing message exceptions. The 5 specialized microprocessor and the host intelligently choose whether a given message or portion of a message is processed by the microprocessor or the host stack.

Brief Description of the Drawings

FIG. 1 is a plan view diagram of a system of the present invention, including a 10 host computer having a communication-processing device for accelerating network communication.

FIG. 2 is a diagram of information flow for the host of FIG. 1 in processing network communication, including a fast-path, a slow-path and a transfer of connection context between the fast and slow-paths.

15 FIG. 3 is a flow chart of message receiving according to the present invention.

FIG. 4 is a diagram of a first embodiment of the host of FIG. 1 including a CPU, a CPD, and a storage device connected by a host bus.

FIG. 5 is a diagram of a second embodiment of the host of FIG. 1 including a CPU, a CPD, and a storage device connected by a host bus.

20 FIG. 6 is a diagram of a third embodiment of the host of FIG. 1 including a CPU, a CPD, and a storage device connected by a host bus.

FIG. 7 is a diagram of a fourth embodiment of the host of FIG. 1 including a CPU, a CPD, and a storage device connected by a host bus.

FIG. 8 is a diagram of a fifth embodiment of the host of FIG. 1 including a CPU, a CPD, and a storage device connected by a host bus.

FIG. 9 is a diagram of a sixth embodiment of the host of FIG. 1 including a CPU, a CPD, and a storage device connected by a host bus.

5 FIG. 10 is a diagram of a seventh embodiment of the host of FIG. 1 including a CPU, a CPD, and a storage device connected by a host bus.

FIG. 11A is a diagram of information flow for the host of FIG. 1 receiving a message packet processed by the slow-path.

10 FIG. 11B is a diagram of information flow for the host of FIG. 1 receiving an initial message packet processed by the fast-path.

FIG. 11C is a diagram of information flow for the host of FIG. 4B receiving a subsequent message packet processed by the fast-path.

FIG. 11D is a diagram of information flow for the host of FIG. 4C receiving a message packet having an error that causes processing to revert to the slow-path.

15 FIG. 12 is a diagram of information flow for the host of FIG. 1 transmitting a message by either the fast or slow-paths.

FIG. 13 is a diagram of information flow for a first embodiment of an intelligent network interface card (INIC) associated with a client having a TCP/IP processing stack.

20 FIG. 14 is a diagram of hardware logic for the INIC embodiment shown in FIG. 13, including a packet control sequencer and a fly-by sequencer.

FIG. 15 is a diagram of the fly-by sequencer of FIG. 14 for analyzing header bytes as they are received by the INIC.

FIG. 16 is a diagram of information flow for a second embodiment of an INIC associated with a server having a TCP/IP processing stack.

FIG. 17 is a diagram of a command driver installed in the host of FIG. 16 for creating and controlling a communication control block for the fast-path.

5 FIG. 18 is a diagram of the TCP/IP stack and command driver of FIG. 17 configured for NetBios communications.

FIG. 19 is a diagram of a communication exchange between the client of FIG. 13 and the server of FIG. 16.

FIG. 20 is a diagram of hardware functions included in the INIC of FIG. 16.

10 FIG. 21 is a diagram of a trio of pipelined microprocessors included in the INIC of FIG. 20, including three phases with a processor in each phase.

FIG. 22A is a diagram of a first phase of the pipelined microprocessor of FIG. 21.

FIG. 22B is a diagram of a second phase of the pipelined microprocessor of FIG. 21.

15 FIG. 22C is a diagram of a third phase of the pipelined microprocessor of FIG. 21.

Detailed Description

FIG. 1 shows a host 20 of the present invention connected by a network 25 to a remote host 22. The increase in processing speed achieved by the present invention can 20 be provided with an intelligent network interface card (INIC) that is easily and affordably added to an existing host, or with a communication processing device (CPD) that is integrated into a host, in either case freeing the host CPU from most protocol processing and allowing improvements in other tasks performed by that CPU. The host 20 in a first

embodiment contains a CPU 28 and a CPD 30 connected by a host bus 33. The CPD 30 includes a microprocessor designed for processing communication data and memory buffers controlled by a direct memory access (DMA) unit. Also connected to the host bus 33 is a storage device 35, such as a semiconductor memory or disk drive, along with 5 any related controls.

Referring additionally to FIG. 2, the host CPU 28 runs a protocol processing stack 44 housed in storage 35, the stack including a data link layer 36, network layer 38, transport layer 40, upper layer 46 and an upper layer interface 42. The upper layer 46 may represent a session, presentation and/or application layer, depending upon the 10 particular protocol being employed and message communicated. The upper layer interface 42, along with the CPU 28 and any related controls, can send or retrieve a file to or from the upper layer 46 or storage 35, as shown by arrow 48. A connection context 50 has been created, as will be explained below, the context summarizing various features of the connection, such as protocol type and source and destination addresses for each 15 protocol layer. The context may be passed between an interface for the session layer 42 and the CPD 30, as shown by arrows 52 and 54, and stored as a communication control block (CCB) at either CPD 30 or storage 35. Control of the CCB can pass between the CPU and the CPD while the CCB remains in storage 35.

When the CPD 30 owns a CCB defining a particular connection, data received by 20 the CPD from the network and pertaining to the connection is referenced to that CCB and can then be sent directly to storage 35 according to a fast-path 58, bypassing sequential protocol processing by the data link 36, network 38 and transport 40 layers. Transmitting a message, such as sending a file from storage 35 to remote host 22, can also occur via

the fast-path 58, in which case the context for the file data is added by the CPD 30 referencing a CCB, rather than by CPU 28 sequentially adding headers during processing by the transport 40, network 38 and data link 36 layers. The DMA controllers of the CPD 30 perform these transfers between CPD and storage 35.

5 The CPD 30 can collapse multiple protocol stacks each having possible separate states into one or more state machines for fast-path processing. As a result, exception conditions may occur that are not provided for in the single state machine, primarily because such conditions occur infrequently and to deal with them on the CPD would provide little or no performance benefit to the host. Such exceptions can be detected by

10 CPD 30 or CPU 28. An advantage of the invention includes the manner in which unexpected situations that occur on a fast-path CCB are handled. The CPD 30 can deal with these rare situations by passing back to the host protocol stack 44 control of the CCB and any associated message frames involved, via a control negotiation. The exception condition is then processed in a conventional manner by the host protocol stack

15 44. At some later time, usually directly after the handling of the exception condition has completed and fast-path processing can resume, the host stack 44 hands control of the CCB back to the CPD.

 This fallback capability enables the performance-impacting functions of the host protocols to be handled by the CPD network microprocessor, while the exceptions are dealt with by the host stacks, the exceptions being so rare as to negligibly effect overall performance. The custom designed network microprocessor can have independent processors for transmitting and receiving network information, and further processors for assisting and queuing, or a single processor can run all the code. One microprocessor

embodiment includes a pipelined trio of receive, transmit and utility processors. Direct memory access (DMA) controllers can be integrated into the CPD to work in close concert with the network microprocessor to quickly move data between buffers within the CPD and other locations such as long term storage. Providing buffers within the CPD 5 reduces host bus traffic by allowing transfers directly to application memory.

FIG. 3 diagrams the general flow of messages received according to the current invention. A large TCP/IP message such as a file transfer may be received by the host from the network in a number of separate, approximately 64 KB transfers, each of which may be split into many, approximately 1.5 KB frames or packets for transmission over a 10 network. Novel NetWare protocol suites running Sequenced Packet Exchange Protocol (SPX) or NetWare Core Protocol (NCP) over Internetwork Packet Exchange (IPX) work in a similar fashion. Another form of data communication which can be handled by the fast-path is Transaction TCP (hereinafter T/TCP or TTCP), a version of TCP which 15 initiates a connection with an initial transaction request after which a reply containing data may be sent according to the connection, rather than initiating a connection via a multi-message initialization dialogue and then transferring data with later messages. In any of the transfers typified by these protocols, each packet conventionally includes a portion of the data being transferred, as well as headers for each of the protocol layers and markers for positioning the packet relative to the rest of the packets of this message.

20 When a message packet or frame is received 47 from a network by the CPD, it is first validated by a hardware assist. This includes determining the protocol types of the various layers, verifying relevant checksums, and summarizing 57 these findings in a status word or words. Included in these words is an indication whether or not the frame

is a candidate for fast-path data flow. Selection 59 of fast-path candidates is based on whether the host may benefit from this message connection being handled by the CPD, which includes determining whether the packet has header bytes denoting particular protocols, such as TCP/IP or SPX/IPX for example. The small percent of frames that are

5 not fast-path candidates are sent 61 to the host protocol stacks for slow-path protocol processing. Subsequent network microprocessor work with each fast-path candidate can determine whether a fast-path connection such as a TCP or SPX CCB is already extant for that candidate, or whether that candidate may be used to set up a new fast-path connection, such as for a TTCP/IP transaction. The validation provided by the CPD

10 provides acceleration regardless of whether a frame is processed by the fast-path or the slow-path, by only providing error free, validated frames to the host CPU for processing.

All received message frames which have been determined by the CPD hardware assist to be fast-path candidates are examined 53 by the network microprocessor or INIC comparator circuits to determine whether they correspond to a CCB controlled by the

15 CPD. Upon confirming such a match, the CPD removes lower layer headers and sends 69 the remaining application data from the frame directly into its final destination in the host using the DMA units of the CPD. This operation may occur immediately upon receipt of a message packet, for example when a TCP connection already exists and destination buffers have been negotiated, or it may first be necessary to process an initial 20 header to acquire a new set of final destination addresses for this transfer. In this latter case, the CPD will queue subsequent message packets while waiting for the destination address, and then DMA the queued application data to that destination.

A fast-path candidate that does not correspond to a CCB may be used to set up a new fast-path connection, by sending 65 the frame to the host for sequential protocol processing. In this case, the host uses this frame to create 51 a CCB, for which ownership which is then passed 67 to the CPD to control subsequent frames on that 5 connection. The CCB, which can be cached in or otherwise controlled by the CPD, includes control and state information pertinent to all protocols that would have been processed had conventional software layer processing been employed. The CCB also contains storage space for per-transfer information used to facilitate moving application-level data contained within subsequent related message packets directly to a host 10 application in a form available for immediate usage. The CPD takes over communication processing for that connection upon receiving control of the CCB from the host.

FIG. 4 shows a view of one embodiment of the host 20, with the CPU 28, the CPD 30 and storage device 35 connected by a host bus 33. The CPU 28 may, for example, be a conventional Pentium® 2, 3 or 4 available from Intel Corporation, 2200 15 Mission College Blvd., Santa Clara, California, 95052. CPU 28 may be integrated with conventional level 1 cache 101 and connected with level 2 cache 102 by a host local bus 103, which is coupled to host bus 33 by a processor interface unit 104, all of which are formed as an integrated circuit 100. Caches 101 and 102 may be composed of static random access memory (SRAM), and more than one such CPU 28 may be included in 20 host 20.

Storage device 35 may include conventional dynamic random access memory (DRAM) 105 and a conventional memory controller 108, which controls writing to and reading from DRAM 105. Although memory 105 is shown in this embodiment as

DRAM, many other known forms of data storage may be used, such as SRAM, ferroelectric random access memory (FRAM) or magnetic random access memory (MRAM).

CPD 30 includes a communications microprocessor 110, microprocessor cache 5 111, SRAM 112, CPD bus 113 and SRAM controller 120. A bus interface unit 121 couples CPD bus 113 and host bus 33. Hardware logic circuits for processing network communication packets include receive and transmit sequencers 114 and 124, receive and transmit media access controllers (MAC) 116 and 126, and physical layer encoder/decoders 118 and 128, respectively. A physical connector, not shown, may 10 couple physical layer encoder/decoder 118 with the network 25, while another physical connector may couple physical layer encoder/decoder 128 with a second network 125.

Although only two networks 25 and 125 are shown connected to host 20 in this figure, CPD may have multiple hardware logic paths supporting multiple physical network connections. Details of hardware logic circuits and a microprocessor that may be 15 included in the CPD are available in the applications mentioned above as being incorporated by reference herein.

A conventional input/output (I/O) controller 131 such as a peripheral component interconnect (PCI) bridge may couple host 20 to a PCI or other I/O bus 133, or the memory controller 108 may instead include an I/O controller. For the case in which I/O 20 controller 131 is provided separately, the memory controller 108 may represent a Graphics and AGP Memory controller Hub (GMCH) portion of an Intel 8xx chipset (e.g., i815, i815E, i820 or i820E), while the PCI bridge or controller 131 may represent the I/O Controller Hub (ICH) portion of that 8xx chipset, available from Intel Corporation. For

the case in which memory controller 108 contains a PCI bridge, memory controller 108 may represent a Northbridge portion of an Intel 440x chipset, part number 440BX, also available, for example, from Via Technologies, Inc., 1045 Mission Court, Fremont CA 94539, as part number VT8501 Apollo MVP4.

5 FIG. 5 shows an embodiment of the host 20 in which the memory controller 108 is integrated with the CPD to form a memory controller and network processor hub 130, for instance on an integrated circuit (IC) chip or printed circuit board (PCB). DRAM 105 may be provided, for example, by plugging into a motherboard slot that connects to the memory controller 108.

10 FIG. 6 shows an embodiment of the host 20 in which the I/O controller 131 is integrated with the CPD to form an I/O controller and network processor hub 140, for instance on an IC or PCB.

15 FIG. 7 shows an embodiment of the host 20 in which the I/O controller 131 and memory controller 108 are integrated with the CPD to form a memory controller, I/O controller and network processor hub 145, for instance on an IC or PCB.

FIG. 8 shows an embodiment of the host 20 in which the CPD 30 is connected to the I/O controller 131, which may for instance be the I/O Controller Hub (ICH) portion of that 8xx chipset (e.g., part number 82801AA ICH or 82801BA ICH2), available from Intel Corporation.

20 FIG. 9 shows an embodiment of the host 20 in which the CPD 30 is connected to a hub interface bus 153 connecting the memory controller 108 and the I/O controller 131. The memory controller 108 may for example be the Graphics and AGP Memory controller Hub (GMCH) portion of an Intel 8xx chipset (e.g., i815, i815E, i820 or i820E),

and the I/O controller 131 may for instance be the I/O Controller Hub (ICH) portion of that 8xx chipset (e.g., part number 82801AA ICH or 82801BA ICH2), available from Intel Corporation.

An advantage of having the CPD 30 integrated into the host 20 occurs when the 5 CPD can access storage 35, such as memory controller 108 and DRAM 105, without using the I/O bus 133. For example, referring again to FIG. 1 and FIGs. 4-10, a logical network connection between host 20 and a remote host 22 may be created by CPU 28 operating a protocol processing stack, the network connection stored as a CCB on DRAM 105. CPD 30 may read the CCB from DRAM 105 and cache the CCB in SRAM 112, 10 also providing a hash table at SRAM 112. A message corresponding to that CCB may then be communicated over network 25 with the assistance of CPD 30 operating a fast-path.

A message corresponding to the CCB and transferred over network 25 from remote host 22 to local host 20, such as a file transferred from computer 22 to computer 15 20 in response to a read request from host 20 to host 22, is sent as a number of packets that are decoded at PHY 118 and accepted by receive MAC unit 116. The sequencer 114 running receive microcode then processes network and transport layer headers of the packets and checksums the packets, creating a descriptor or summary of each packet that is buffered in SRAM 112 and then written with the packet to DRAM 105 by DMA 115.

20 An initial packet of the message may contain a session layer header or other information describing the message, and that information may be processed by the CPU running a file system or other management system to obtain a destination for the data from the message (e.g., a portion of DRAM 105 earmarked by the file system for holding

all the message data). Once this destination is provided, memory controller 108 moves the data from each message packet, without the headers or descriptors, from the initial storage location in DRAM 105 to the destination, to be used for example by an application. This data movement can be accomplished by either a DMA or remapping of 5 memory blocks on the DRAM 105 by memory controller 108.

Alternatively, after processing by CPD hardware logic of the packets as they arrive from the network, the packets and their corresponding descriptors or summaries may be stored in SRAM 112 or other memory such as DRAM, not shown in this figure, controlled by the CPD. Then, after a message or session layer header has been processed 10 and the destination for the message has been assigned in DRAM 105, DMA units of CPD work with memory controller 108 to move the data from each message packet, without the headers or descriptors, from the original storage location in SRAM 112 to the destination, to be available for example to an application.

The CCB may likewise be cached in SRAM 112 or other memory such as DRAM 15 controlled by the CPD, as well as in DRAM 105. With the CCB stored in DRAM 105, either the communication microprocessor 110 or CPU 28 can access the CCB. For example, after creation of the CCB by CPU 28 running the protocol stack 44, the stack 20 can issue a command to the CPD 30 to take ownership of the CCB, accompanied by a list of addresses in DRAM 105 of where the CCB is stored. A response by the CPD accepting control of the CCB accomplishes the ownership change, and the CPD may then access the CCB as needed. The CCB may contain an ownership bit that designates whether the CPU 28 or CPD 30 controls the CCB. Other mechanisms of allocating control of the CCB stored in DRAM 105 are also possible.

Alternatively, the CCB can be copied from DRAM 105 into SRAM 112 or other memory such as DRAM controlled by the CPD 30. With CCB cached in SRAM 112, each received packet corresponding to the CCB can be stored, along with its descriptor, in DRAM 105 directly after processing by the CPD hardware logic. For the case in 5 which a destination for the message being received has already been allocated, for example by processing a session layer header for the message, data from the received packet can be stored in the destination in DRAM 105 directly after processing by the CPD hardware logic. Instead of caching a complete CCB in SRAM 112, selected cache lines of the CCB may be cached in microprocessor cache 111 by a cache manager run by 10 communication microprocessor 110, and controlling DMA 115.

In one embodiment, the communication microprocessor 110 can create a CCB by processing connection initiation communications for selected fast-path protocols, such as TCP/IP, by running a protocol stack stored in SRAM 112 or other memory such as DRAM controlled by the CPD 30, with the CPU 28 available for processing other 15 protocols. In this case, a source or destination for a message corresponding to the CCB can be provided by the host file system, by respectively creating or analyzing information from a session layer or other header for the message.

FIG. 10 shows an embodiment similar to that shown in FIG. 4, but with the communication microprocessor 110 directly coupled to host bus 33 via the bus interface unit 121. The communication microprocessor 110 in the embodiments illustrated in 20 FIGs. 5-9 can also be placed on the host bus in those embodiments. Locating the communication microprocessor 110 on the host bus in this fashion allows the communication microprocessor 110 to directly access CCBs stored in DRAM 105, rather

than directing DMA controllers to perform that function. Further, conventional processors such as an Intel i960 may be employed for communication microprocessor 110 in the embodiments for which the communication microprocessor 110 is disposed on host bus 33.

5 As shown more specifically in FIG. 11A, when a message packet is received from the remote host 22 via network 25, the packet enters hardware receive logic 32 of the CPD 30, which checksums headers and data, and parses the headers, creating a word or words which identify the message packet and status, storing the headers, data and word temporarily in memory 60. As well as validating the packet, the receive logic 32
10 indicates with the word whether this packet is a candidate for fast-path processing. FIG. 11A depicts the case in which the packet is not a fast-path candidate, in which case the CPD 30 sends the validated headers and data from memory 60 to data link layer 36 along an internal bus for processing by the host CPU, as shown by arrow 56. The packet is processed by the host protocol stack 44 of data link 36, network 38, transport 40 and
15 session 42 layers, and data (D) 63 from the packet may then be sent to storage 35, as shown by arrow 65.

FIG. 11B, depicts the case in which the receive logic 32 of the CPD determines that a message packet is a candidate for fast-path processing, for example by deriving from the packet's headers that the packet belongs to a TCP/IP, TTCP/IP or SPX/IPX
20 message. A processor 55 in the CPD 30 then checks to see whether the word that summarizes the fast-path candidate matches a CCB held in a cache 62. Upon finding no match for this packet, the CPD sends the validated packet from memory 60 to the host protocol stack 44 for processing. Host stack 44 may use this packet to create a

connection context for the message, including finding and reserving a destination for data from the message associated with the packet, the context taking the form of a CCB. The present embodiment employs a single specialized host stack 44 for processing both fast-path and non-fast-path candidates, while in an embodiment described below fast-path 5 candidates are processed by a different host stack than non-fast-path candidates. Some data (D1) 66 from that initial packet may optionally be sent to the destination in storage 35, as shown by arrow 68. The CCB is then sent to the CPD 30 to be saved in cache 62, as shown by arrow 64. For a traditional connection-based message such as typified by TCP/IP, the initial packet may be part of a connection initialization dialogue that 10 transpires between hosts before the CCB is created and passed to the CPD 30.

Referring now to FIG. 11C, when a subsequent packet from the same connection as the initial packet is received from the network 25 by CPD 30, the packet headers and data are validated by the receive logic 32, and the headers are parsed to create a summary of the message packet and a hash for finding a corresponding CCB, the summary and 15 hash contained in a word or words. The word or words are temporarily stored in memory 60 along with the packet. The processor 55 checks for a match between the hash and each CCB that is stored in the cache 62 and, finding a match, sends the data (D2) 70 via a fast-path directly to the destination in storage 35, as shown by arrow 72, bypassing the session layer 42, transport layer 40, network layer 38 and data link layer 36. The 20 remaining data packets from the message can also be sent by DMA directly to storage, avoiding the relatively slow protocol layer processing and repeated copying by the CPU stack 44.

FIG. 11D shows the procedure for handling the rare instance when a message for which a fast-path connection has been established, such as shown in FIG. 11C, has a packet that is not easily handled by the CPD. In this case the packet is sent to be processed by the protocol stack 44, which is handed the CCB for that message from 5 cache 62 via a control dialogue with the CPD, as shown by arrow 76, signaling to the CPU to take over processing of that message. Slow-path processing by the protocol stack then results in data (D3) 80 from the packet being sent, as shown by arrow 82, to storage 35. Once the packet has been processed and the error situation corrected, the CCB can be handed back via a control dialogue to the cache 62, so that payload data from subsequent 10 packets of that message can again be sent via the fast-path of the CPD 30. Thus the CPU and CPD together decide whether a given message is to be processed according to fast-path hardware processing or more conventional software processing by the CPU.

Transmission of a message from the host 20 to the network 25 for delivery to remote host 22 also can be processed by either sequential protocol software processing 15 via the CPU or accelerated hardware processing via the CPD 30, as shown in FIG. 12. A message (M) 90 that is selected by CPU 28 from storage 35 can be sent to session layer 42 for processing by stack 44, as shown by arrows 92 and 96. For the situation in which a connection exists and the CPD 30 already has an appropriate CCB for the message, however, data packets can bypass host stack 44 and be sent by DMA directly to memory 20 60, with the processor 55 adding to each data packet a single header containing all the appropriate protocol layers, and sending the resulting packets to the network 25 for transmission to remote host 22. This fast-path transmission can greatly accelerate processing for even a single packet, with the acceleration multiplied for a larger message.

A message for which a fast-path connection is not extant thus may benefit from creation of a CCB with appropriate control and state information for guiding fast-path transmission. For a traditional connection-based message, such as typified by TCP/IP or SPX/IPX, the CCB is created during connection initialization dialogue. For a quick-
5 connection message, such as typified by TTCP/IP, the CCB can be created with the same transaction that transmits payload data. In this case, the transmission of payload data may be a reply to a request that was used to set up the fast-path connection. In any case, the CCB provides protocol and status information regarding each of the protocol layers, including which user is involved and storage space for per-transfer information. The
10 CCB is created by protocol stack 44, which then passes the CCB to the CPD 30 by writing to a command register of the CPD, as shown by arrow 98. Guided by the CCB, the processor 55 moves network frame-sized portions of the data from the source in host memory 35 into its own memory 60 using DMA, as depicted by arrow 99. The processor 55 then prepends appropriate headers and checksums to the data portions, and transmits
15 the resulting frames to the network 25, consistent with the restrictions of the associated protocols. After the CPD 30 has received an acknowledgement that all the data has reached its destination, the CPD will then notify the host 35 by writing to a response buffer.

Thus, fast-path transmission of data communications also relieves the host CPU
20 of per-frame processing. A vast majority of data transmissions can be sent to the network by the fast-path. Both the input and output fast-paths attain a huge reduction in interrupts by functioning at an upper layer level, i.e., session level or higher, and interactions between the network microprocessor and the host occur using the full transfer sizes

which that upper layer wishes to make. For fast-path communications, an interrupt only occurs (at the most) at the beginning and end of an entire upper-layer message transaction, and there are no interrupts for the sending or receiving of each lower layer portion or packet of that transaction.

5 A simplified intelligent network interface card (INIC) 150 is shown in FIG. 12 to provide a network interface for a host 152. Hardware logic 171 of the INIC 150 is connected to a network 155, with a peripheral bus (PCI) 157 connecting the INIC and host. The host 152 in this embodiment has a TCP/IP protocol stack, which provides a slow-path 158 for sequential software processing of message frames received from the 10 network 155. The host 152 protocol stack includes a data link layer 160, network layer 162, a transport layer 164 and an application layer 166, which provides a source or destination 168 for the communication data in the host 152. Other layers which are not shown, such as session and presentation layers, may also be included in the host stack 152, and the source or destination may vary depending upon the nature of the data and 15 may actually be the application layer.

The INIC 150 has a network processor 170 which chooses between processing messages along a slow-path 158 that includes the protocol stack of the host, or along a fast-path 159 that bypasses the protocol stack of the host. Each received packet is processed on the fly by hardware logic 171 contained in INIC 150, so that all of the 20 protocol headers for a packet can be processed without copying, moving or storing the data between protocol layers. The hardware logic 171 processes the headers of a given packet at one time as packet bytes pass through the hardware, by categorizing selected header bytes. Results of processing the selected bytes help to determine which other

bytes of the packet are categorized, until a summary of the packet has been created, including checksum validations. The processed headers and data from the received packet are then stored in INIC storage 185, as well as the word or words summarizing the headers and status of the packet.

- 5 The hardware processing of message packets received by INIC 150 from network 155 is shown in more detail in FIG. 14. A received message packet first enters a media access controller 172, which controls INIC access to the network and receipt of packets and can provide statistical information for network protocol management. From there, data flows one byte at a time into an assembly register 174, which in this example is 128
- 10 bits wide. The data is categorized by a fly-by sequencer 178, as will be explained in more detail with regard to FIG. 15, which examines the bytes of a packet as they fly by, and generates status from those bytes that will be used to summarize the packet. The status thus created is merged with the data by a multiplexor 180 and the resulting data stored in SRAM 182. A packet control sequencer 176 oversees the fly-by sequencer 178,
- 15 examines information from the media access controller 172, counts the bytes of data, generates addresses, moves status and manages the movement of data from the assembly register 174 to SRAM 182 and eventually DRAM 188. The packet control sequencer 176 manages a buffer in SRAM 182 via SRAM controller 183, and also indicates to a DRAM controller 186 when data needs to be moved from SRAM 182 to a buffer in DRAM 188.
- 20 Once data movement for the packet has been completed and all the data has been moved to the buffer in DRAM 188, the packet control sequencer 176 will move the status that has been generated in the fly-by sequencer 178 out to the SRAM 182 and to the beginning of the DRAM 188 buffer to be prepended to the packet data. The packet

control sequencer 176 then requests a queue manager 184 to enter a receive buffer descriptor into a receive queue, which in turn notifies the processor 170 that the packet has been processed by hardware logic 171 and its status summarized.

FIG. 15 shows that the fly-by sequencer 178 has several tiers, with each tier 5 generally focusing on a particular portion of the packet header and thus on a particular protocol layer, for generating status pertaining to that layer. The fly-by sequencer 178 in this embodiment includes a media access control sequencer 191, a network sequencer 192, a transport sequencer 194 and a session sequencer 195. Sequencers pertaining to higher protocol layers can additionally be provided. The fly-by sequencer 178 is reset by 10 the packet control sequencer 176 and given pointers by the packet control sequencer that tell the fly-by sequencer whether a given byte is available from the assembly register 174. The media access control sequencer 191 determines, by looking at bytes 0-5, that a packet is addressed to host 152 rather than or in addition to another host. Offsets 12 and 15 13 of the packet are also processed by the media access control sequencer 191 to determine the type field, for example whether the packet is Ethernet or 802.3. If the type field is Ethernet those bytes also tell the media access control sequencer 191 the packet's network protocol type. For the 802.3 case, those bytes instead indicate the length of the entire frame, and the media access control sequencer 191 will check eight bytes further into the packet to determine the network layer type.

20 For most packets the network sequencer 192 validates that the header length received has the correct length, and checksums the network layer header. For fast-path candidates the network layer header is known to be IP or IPX from analysis done by the media access control sequencer 191. Assuming for example that the type field is 802.3

and the network protocol is IP, the network sequencer 192 analyzes the first bytes of the network layer header, which will begin at byte 22, in order to determine IP type. The first bytes of the IP header will be processed by the network sequencer 192 to determine what IP type the packet involves. Determining that the packet involves, for example, IP version 4, directs further processing by the network sequencer 192, which also looks at the protocol type located ten bytes into the IP header for an indication of the transport header protocol of the packet. For example, for IP over Ethernet, the IP header begins at offset 14, and the protocol type byte is offset 23, which will be processed by network logic to determine whether the transport layer protocol is TCP, for example. From the length of the network layer header, which is typically 20-40 bytes, network sequencer 192 determines the beginning of the packet's transport layer header for validating the transport layer header. Transport sequencer 194 may generate checksums for the transport layer header and data, which may include information from the IP header in the case of TCP at least.

Continuing with the example of a TCP packet, transport sequencer 194 also analyzes the first few bytes in the transport layer portion of the header to determine, in part, the TCP source and destination ports for the message, such as whether the packet is NetBios or other protocols. Byte 12 of the TCP header is processed by the transport sequencer 194 to determine and validate the TCP header length. Byte 13 of the TCP header contains flags that may, aside from ack flags and push flags, indicate unexpected options, such as reset and fin, that may cause the processor to categorize this packet as an exception. TCP offset bytes 16 and 17 are the checksum, which is pulled out and stored by the hardware logic 171 while the rest of the frame is validated against the checksum.

Session sequencer 195 determines the length of the session layer header, which in the case of NetBios is only four bytes, two of which tell the length of the NetBios payload data, but which can be much larger for other protocols. The session sequencer 195 can also be used to categorize the type of message as read or write, for example, for 5 which the fast-path may be particularly beneficial. Further upper layer logic processing, depending upon the message type, can be performed by the hardware logic 171 of packet control sequencer 176 and fly-by sequencer 178. Thus hardware logic 171 intelligently directs hardware processing of the headers by categorization of selected bytes from a single stream of bytes, with the status of the packet being built from classifications 10 determined on the fly. Once the packet control sequencer 176 detects that all of the packet has been processed by the fly-by sequencer 178, the packet control sequencer 176 adds the status information generated by the fly-by sequencer 178 and any status information generated by the packet control sequencer 176, and prepends (adds to the front) that status information to the packet, for convenience in handling the packet by the 15 processor 170. The additional status information generated by the packet control sequencer 176 includes media access controller 172 status information and any errors discovered, or data overflow in either the assembly register or DRAM buffer, or other miscellaneous information regarding the packet. The packet control sequencer 176 also stores entries into a receive buffer queue and a receive statistics queue via the queue 20 manager 184.

An advantage of processing a packet by hardware logic 171 is that the packet does not, in contrast with conventional sequential software protocol processing, have to be stored, moved, copied or pulled from storage for processing each protocol layer header,

offering dramatic increases in processing efficiency and savings in processing time for each packet. The packets can be processed at the rate bits are received from the network, for example 100 megabits/second for a 100 baseT connection. The time for categorizing a packet received at this rate and having a length of sixty bytes is thus about 5

5 microseconds. The total time for processing this packet with the hardware logic 171 and sending packet data to its host destination via the fast-path may be about 16 microseconds or less, assuming a 66 MH PCI bus, whereas conventional software protocol processing by a 300 MH Pentium II[®] processor may take as much as 200 microseconds in a busy system. More than an order of magnitude decrease in processing time can thus be
10 achieved with fast-path 159 in comparison with a high-speed CPU employing conventional sequential software protocol processing, demonstrating the dramatic acceleration provided by processing the protocol headers by the hardware logic 171 and processor 170, without even considering the additional time savings afforded by the reduction in CPU interrupts and host bus bandwidth savings.

15 The processor 170 chooses, for each received message packet held in storage 185, whether that packet is a candidate for the fast-path 159 and, if so, checks to see whether a fast-path has already been set up for the connection that the packet belongs to. To do this, the processor 170 first checks the header status summary to determine whether the packet headers are of a protocol defined for fast-path candidates. If not, the
20 processor 170 commands DMA controllers in the INIC 150 to send the packet to the host for slow-path 158 processing. Even for a slow-path 158 processing of a message, the INIC 150 thus performs initial procedures such as validation and determination of

message type, and passes the validated message at least to the data link layer 160 of the host.

For fast-path 159 candidates, the processor 170 checks to see whether the header status summary matches a CCB held by the INIC. If so, the data from the packet is sent

5 along fast-path 159 to the destination 168 in the host. If the fast-path 159 candidate's packet summary does not match a CCB held by the INIC, the packet may be sent to the host 152 for slow-path processing to create a CCB for the message. Employment of the fast-path 159 may also not be needed or desirable for the case of fragmented messages or other complexities. For the vast majority of messages, however, the INIC fast-path 159

10 can greatly accelerate message processing. The INIC 150 thus provides a single state machine processor 170 that decides whether to send data directly to its destination, based upon information gleaned on the fly, as opposed to the conventional employment of a state machine in each of several protocol layers for determining the destiny of a given packet.

15 In processing an indication or packet received at the host 152, a protocol driver of the host selects the processing route based upon whether the indication is fast-path or slow-path. A TCP/IP or SPX/IPX message has a connection that is set up from which a CCB is formed by the driver and passed to the INIC for matching with and guiding the fast-path packet to the connection destination 168. For a TTCP/IP message, the driver

20 can create a connection context for the transaction from processing an initial request packet, including locating the message destination 168, and then passing that context to the INIC in the form of a CCB for providing a fast-path for a reply from that destination.

A CCB includes connection and state information regarding the protocol layers and

5 packets of the message. Thus a CCB can include source and destination media access control (MAC) addresses, source and destination IP or IPX addresses, source and destination TCP or SPX ports, TCP variables such as timers, receive and transmit windows for sliding window protocols, and information denoting the session layer protocol.

10 Caching the CCBs in a hash table in the INIC provides quick comparisons with words summarizing incoming packets to determine whether the packets can be processed via the fast-path 159, while the full CCBs are also held in the INIC for processing. Other ways to accelerate this comparison include software processes such as a B-tree or hardware assists such as a content addressable memory (CAM). When INIC microcode or comparitor circuits detect a match with the CCB, a DMA controller places the data from the packet in the destination 168, without any interrupt by the CPU, protocol processing or copying. Depending upon the type of message received, the destination of the data may be the session, presentation or application layers, or a file buffer cache in 15 the host 152.

20 FIG. 16 shows an INIC 200 connected to a host 202 that is employed as a file server. This INIC provides a network interface for several network connections employing the 802.3u standard, commonly known as Fast Ethernet. The INIC 200 is connected by a PCI bus 205 to the server 202, which maintains a TCP/IP or SPX/IPX protocol stack including MAC layer 212, network layer 215, transport layer 217 and application layer 220, with a source/destination 222 shown above the application layer, although as mentioned earlier the application layer can be the source or destination. The INIC is also connected to network lines 210, 240, 242 and 244, which are preferably fast

Ethernet, twisted pair, fiber optic, coaxial cable or other lines each allowing data transmission of 100 Mb/s, while faster and slower data rates are also possible. Network lines 210, 240, 242 and 244 are each connected to a dedicated row of hardware circuits which can each validate and summarize message packets received from their respective 5 network line. Thus line 210 is connected with a first horizontal row of sequencers 250, line 240 is connected with a second horizontal row of sequencers 260, line 242 is connected with a third horizontal row of sequencers 262 and line 244 is connected with a fourth horizontal row of sequencers 264. After a packet has been validated and summarized by one of the horizontal hardware rows it is stored along with its status 10 summary in storage 270.

A network processor 230 determines, based on that summary and a comparison with any CCBs stored in the INIC 200, whether to send a packet along a slow-path 231 for processing by the host. A large majority of packets can avoid such sequential processing and have their data portions sent by DMA along a fast-path 237 directly to the 15 data destination 222 in the server according to a matching CCB. Similarly, the fast-path 237 provides an avenue to send data directly from the source 222 to any of the network lines by processor 230 division of the data into packets and addition of full headers for network transmission, again minimizing CPU processing and interrupts. For clarity only horizontal sequencer 250 is shown active; in actuality each of the sequencer rows 250, 260, 262 and 264 offers full duplex communication, concurrently with all other sequencer rows. The specialized INIC 200 is much faster at working with message packets than 20 even advanced general-purpose host CPUs that processes those headers sequentially according to the software protocol stack.

One of the most commonly used network protocols for large messages such as file transfers is server message block (SMB) over TCP/IP. SMB can operate in conjunction with redirector software that determines whether a required resource for a particular operation, such as a printer or a disk upon which a file is to be written, resides in or is 5 associated with the host from which the operation was generated or is located at another host connected to the network, such as a file server. SMB and server/redirector are conventionally serviced by the transport layer; in the present invention SMB and redirector can instead be serviced by the INIC. In this case, sending data by the DMA controllers from the INIC buffers when receiving a large SMB transaction may greatly 10 reduce interrupts that the host must handle. Moreover, this DMA generally moves the data to its final destination in the file system cache. An SMB transmission of the present invention follows essentially the reverse of the above described SMB receive, with data transferred from the host to the INIC and stored in buffers, while the associated protocol headers are prepended to the data in the INIC, for transmission via a network line to a 15 remote host. Processing by the INIC of the multiple packets and multiple TCP, IP, NetBios and SMB protocol layers via custom hardware and without repeated interrupts of the host can greatly increase the speed of transmitting an SMB message to a network line.

As shown in FIG. 17, for controlling whether a given message is processed by the host 202 or by the INIC 200, a message command driver 300 may be installed in host 202 20 to work in concert with a host protocol stack 310. The command driver 300 can intervene in message reception or transmittal, create CCBs and send or receive CCBs from the INIC 200, so that functioning of the INIC, aside from improved performance, is transparent to a user. Also shown is an INIC memory 304 and an INIC miniport driver

306, which can direct message packets received from network 210 to either the conventional protocol stack 310 or the command protocol stack 300, depending upon whether a packet has been labeled as a fast-path candidate. The conventional protocol stack 310 has a data link layer 312, a network layer 314 and a transport layer 316 for 5 conventional, lower layer processing of messages that are not labeled as fast-path candidates and therefore not processed by the command stack 300. Residing above the lower layer stack 310 is an upper layer 318, which represents a session, presentation and/or application layer, depending upon the message communicated. The command driver 300 similarly has a data link layer 320, a network layer 322 and a transport layer 10 325.

The driver 300 includes an upper layer interface 330 that determines, for transmission of messages to the network 210, whether a message transmitted from the upper layer 318 is to be processed by the command stack 300 and subsequently the INIC fast-path, or by the conventional stack 310. When the upper layer interface 330 receives 15 an appropriate message from the upper layer 318 that would conventionally be intended for transmission to the network after protocol processing by the protocol stack of the host, the message is passed to driver 300. The INIC then acquires network-sized portions of the message data for that transmission via INIC DMA units, prepends headers to the data portions and sends the resulting message packets down the wire. Conversely, in 20 receiving a TCP, TTCP, SPX or similar message packet from the network 210 to be used in setting up a fast-path connection, miniport driver 306 diverts that message packet to command driver 300 for processing. The driver 300 processes the message packet to create a context for that message, with the driver 302 passing the context and command

instructions back to the INIC 200 as a CCB for sending data of subsequent messages for the same connection along a fast-path. Hundreds of TCP, TTCP, SPX or similar CCB connections may be held indefinitely by the INIC, although a least recently used (LRU) algorithm is employed for the case when the INIC cache is full. The driver 300 can also 5 create a connection context for a TTCP request which is passed to the INIC 200 as a CCB, allowing fast-path transmission of a TTCP reply to the request. A message having a protocol that is not accelerated can be processed conventionally by protocol stack 310.

FIG. 18 shows a TCP/IP implementation of command driver software for Microsoft® protocol messages. A conventional host protocol stack 350 includes MAC 10 layer 353, IP layer 355 and TCP layer 358. A command driver 360 works in concert with the host stack 350 to process network messages. The command driver 360 includes a MAC layer 363, an IP layer 366 and an Alacritech TCP (ATCP) layer 373. The conventional stack 350 and command driver 360 share a network driver interface specification (NDIS) layer 375, which interacts with the INIC miniport driver 306. The 15 INIC miniport driver 306 sorts receive indications for processing by either the conventional host stack 350 or the ATCP driver 360. A TDI filter driver and upper layer interface 380 similarly determines whether messages sent from a TDI user 382 to the network are diverted to the command driver and perhaps to the fast-path of the INIC, or processed by the host stack.

20 FIG. 19 depicts a typical SMB exchange between a client 190 and server 290, both of which have communication devices of the present invention, the communication devices each holding a CCB defining their connection for fast-path movement of data. The client 190 includes INIC 150, 802.3 compliant data link layer 160, IP layer 162, TCP

layer 164, NetBios layer 166, and SMB layer 168. The client has a slow-path 157 and fast-path 159 for communication processing. Similarly, the server 290 includes INIC 200, 802.3 compliant data link layer 212, IP layer 215, TCP layer 217, NetBios layer 220, and SMB 222. The server is connected to network lines 240, 242 and 244, as well as line 5 210 which is connected to client 190. The server also has a slow-path 231 and fast-path 237 for communication processing.

Assuming that the client 190 wishes to read a 100KB file on the server 290, the client may begin by sending a Read Block Raw (RBR) SMB command across network 210 requesting the first 64 KB of that file on the server 290. The RBR command may be 10 only 76 bytes, for example, so the INIC 200 on the server will recognize the message type (SMB) and relatively small message size, and send the 76 bytes directly via the fast-path to NetBios of the server. NetBios will give the data to SMB, which processes the Read request and fetches the 64KB of data into server data buffers. SMB then calls NetBios to send the data, and NetBios outputs the data for the client. In a conventional 15 host, NetBios would call TCP output and pass 64 KB to TCP, which would divide the data into 1460 byte segments and output each segment via IP and eventually MAC (slow-path 231). In the present case, the 64KB data goes to the ATCP driver along with an indication regarding the client-server SMB connection, which denotes a CCB held by the INIC. The INIC 200 then proceeds to DMA 1460 byte segments from the host buffers, 20 add the appropriate headers for TCP, IP and MAC at one time, and send the completed packets on the network 210 (fast-path 237). The INIC 200 will repeat this until the whole 64KB transfer has been sent. Usually after receiving acknowledgement from the client

that the 64KB has been received, the INIC will then send the remaining 36KB also by the fast-path 237.

With INIC 150 operating on the client 190 when this reply arrives, the INIC 150 recognizes from the first frame received that this connection is receiving fast-path 159 processing (TCP/IP, NetBios, matching a CCB), and the ATCP may use this first frame to acquire buffer space for the message. This latter case is done by passing the first 128 bytes of the NetBios portion of the frame via the ATCP fast-path directly to the host NetBios; that will give NetBios/SMB all of the frame's headers. NetBios/SMB will analyze these headers, realize by matching with a request ID that this is a reply to the original RawRead connection, and give the ATCP a 64K list of buffers into which to place the data. At this stage only one frame has arrived, although more may arrive while this processing is occurring. As soon as the client buffer list is given to the ATCP, it passes that transfer information to the INIC 150, and the INIC 150 starts DMAing any frame data that has accumulated into those buffers.

FIG. 20 provides a simplified diagram of the INIC 200, which combines the functions of a network interface controller and a protocol processor in a single ASIC chip 400. The INIC 200 in this embodiment offers a full-duplex, four channel, 10/100-Megabit per second (Mbps) intelligent network interface controller that is designed for high speed protocol processing for server applications. Although designed specifically for server applications, the INIC 200 can be connected to personal computers, workstations, routers or other hosts anywhere that TCP/IP, TTCP/IP or SPX/IPX protocols are being utilized.

The INIC 200 is connected with four network lines 210, 240, 242 and 244, which may transport data along a number of different conduits, such as twisted pair, coaxial cable or optical fiber, each of the connections providing a media independent interface (MII). The lines preferably are 802.3 compliant and in connection with the INIC 5 constitute four complete Ethernet nodes, the INIC supporting 10Base-T, 10Base-T2, 100Base-TX, 100Base-FX and 100Base-T4 as well as future interface standards. Physical layer identification and initialization is accomplished through host driver initialization routines. The connection between the network lines 210, 240, 242 and 244 and the INIC 200 is controlled by MAC units MAC-A 402, MAC-B 404, MAC-C 406 10 and MAC-D 408 which contain logic circuits for performing the basic functions of the MAC sublayer, essentially controlling when the INIC accesses the network lines 210, 240, 242 and 244. The MAC units 402-408 may act in promiscuous, multicast or unicast modes, allowing the INIC to function as a network monitor, receive broadcast and multicast packets and implement multiple MAC addresses for each node. The MAC 15 units 402-408 also provide statistical information that can be used for simple network management protocol (SNMP).

The MAC units 402, 404, 406 and 408 are each connected to a transmit and receive sequencer, XMT & RCV-A 418, XMT & RCV-B 420, XMT & RCV-C 422 and XMT & RCV-D 424, by wires 410, 412, 414 and 416, respectively. Each of the transmit 20 and receive sequencers can perform several protocol processing steps on the fly as message frames pass through that sequencer. In combination with the MAC units, the transmit and receive sequencers 418-422 can compile the packet status for the data link, network, transport, session and, if appropriate, presentation and application layer

protocols in hardware, greatly reducing the time for such protocol processing compared to conventional sequential software engines. The transmit and receive sequencers 410-414 are connected, by lines 426, 428, 430 and 432 to an SRAM and DMA controller 444, which includes DMA controllers 438 and SRAM controller 442. Static random access 5 memory (SRAM) buffers 440 are coupled with SRAM controller 442 by line 441. The SRAM and DMA controllers 444 interact across line 446 with external memory control 450 to send and receive frames via external memory bus 455 to and from dynamic random access memory (DRAM) buffers 460, which is located adjacent to the IC chip 400. The DRAM buffers 460 may be configured as 4 MB, 8 MB, 16 MB or 32 MB, and 10 may optionally be disposed on the chip. The SRAM and DMA controllers 444 are connected via line 464 to a PCI Bus Interface Unit (BIU) 468, which manages the interface between the INIC 200 and the PCI interface bus 257. The 64-bit, multiplexed BIU 380 provides a direct interface to the PCI bus 257 for both slave and master functions. The INIC 200 is capable of operating in either a 64-bit or 32-bit PCI 15 environment, while supporting 64-bit addressing in either configuration.

A microprocessor 470 is connected by line 472 to the SRAM and DMA controllers 444, and connected via line 475 to the PCI BIU 468. Microprocessor 470 instructions and register files reside in an on chip control store 480, which includes a writable on-chip control store (WCS) of SRAM and a read only memory (ROM), and is 20 connected to the microprocessor by line 477. The microprocessor 470 offers a programmable state machine which is capable of processing incoming frames, processing host commands, directing network traffic and directing PCI bus traffic. Three processors are implemented using shared hardware in a three level pipelined architecture that

launches and completes a single instruction for every clock cycle. A receive processor 482 is dedicated to receiving communications while a transmit processor 484 is dedicated to transmitting communications in order to facilitate full duplex communication, while a utility processor 486 offers various functions including overseeing and controlling PCI 5 register access. The instructions for the three processors 482, 484 and 486 reside in the on-chip control-store 480.

The INIC 200 in this embodiment can support up to 256 CCBs which are maintained in a table in the DRAM 460. There is also, however, a CCB index in hash order in the SRAM 440 to save sequential searching. Once a hash has been generated, 10 the CCB is cached in SRAM, with up to sixteen cached CCBs in SRAM in this example. These cache locations are shared between the transmit 484 and receive 486 processors so that the processor with the heavier load is able to use more cache buffers. There are also eight header buffers and eight command buffers to be shared between the sequencers. A given header or command buffer is not statically linked to a specific CCB buffer, as the 15 link is dynamic on a per-frame basis.

FIG. 21 shows an overview of the pipelined microprocessor 470, in which instructions for the receive, transmit and utility processors are executed in three distinct phases according to Clock increments I, II and III, the phases corresponding to each of the pipeline stages. Each phase is responsible for different functions, and each of the 20 three processors occupies a different phase during each Clock increment. Each processor usually operates upon a different instruction stream from the control store 480, and each carries its own program counter and status through each of the phases.

In general, a first instruction phase 500 of the pipelined microprocessors completes an instruction and stores the result in a destination operand, fetches the next instruction, and stores that next instruction in an instruction register. A first register set 490 provides a number of registers including the instruction register, and a set of controls 492 for first register set provides the controls for storage to the first register set 490.

5 Some items pass through the first phase without modification by the controls 492, and instead are simply copied into the first register set 490 or a RAM file register 533. A second instruction phase 560 has an instruction decoder and operand multiplexer 498 that generally decodes the instruction that was stored in the instruction register of the first register set 490 and gathers any operands which have been generated, which are then stored in a decode register of a second register set 496. The first register set 490, second register set 496 and a third register set 501, which is employed in a third instruction phase 600, include many of the same registers, as will be seen in the more detailed views of FIGs. 14 A-C. The instruction decoder and operand multiplexer 498 can read from two address and data ports of the RAM file register 533, which operates in both the first phase 500 and second phase 560. A third phase 600 of the processor 470 has an arithmetic logic unit (ALU) 602 which generally performs any ALU operations on the operands from the second register set, storing the results in a results register included in the third register set 501. A stack exchange 608 can reorder register stacks, and a queue manager 20 503 can arrange queues for the processor 470, the results of which are stored in the third register set.

The instructions continue with the first phase then following the third phase, as depicted by a circular pipeline 505. Note that various functions have been distributed

across the three phases of the instruction execution in order to minimize the combinatorial delays within any given phase. With a frequency in this embodiment of 66 Megahertz, each Clock increment takes 15 nanoseconds to complete, for a total of 45 nanoseconds to complete one instruction for each of the three processors. The instruction 5 phases are depicted in more detail in FIGs. 15A-C, in which each phase is shown in a different figure.

More particularly, FIG. 22A shows some specific hardware functions of the first phase 500, which generally includes the first register set 490 and related controls 492. The controls for the first register set 492 includes an SRAM control 502, which is a 10 logical control for loading address and write data into SRAM address and data registers 520. Thus the output of the ALU 602 from the third phase 600 may be placed by SRAM control 502 into an address register or data register of SRAM address and data registers 520. A load control 504 similarly provides controls for writing a context for a file to file context register 522, and another load control 506 provides controls for storing a variety 15 of miscellaneous data to flip-flop registers 525. ALU condition codes, such as whether a carried bit is set, get clocked into ALU condition codes register 528 without an operation performed in the first phase 500. Flag decodes 508 can perform various functions, such as setting locks, that get stored in flag registers 530.

The RAM file register 533 has a single write port for addresses and data and two 20 read ports for addresses and data, so that more than one register can be read from at one time. As noted above, the RAM file register 533 essentially straddles the first and second phases, as it is written in the first phase 500 and read from in the second phase 560. A control store instruction 510 allows the reprogramming of the processors due to new data

in from the control store 480, not shown in this figure, the instructions stored in an instruction register 535. The address for this is generated in a fetch control register 511, which determines which address to fetch, the address stored in fetch address register 538. Load control 515 provides instructions for a program counter 540, which operates much 5 like the fetch address for the control store. A last-in first-out stack 544 of three registers is copied to the first register set without undergoing other operations in this phase. Finally, a load control 517 for a debug address 548 is optionally included, which allows correction of errors that may occur.

FIG. 22B depicts the second microprocessor phase 560, which includes reading 10 addresses and data out of the RAM file register 533. A scratch SRAM 565 is written from SRAM address and data register 520 of the first register set, which includes a register that passes through the first two phases to be incremented in the third. The scratch SRAM 565 is read by the instruction decoder and operand multiplexer 498, as are most of the registers from the first register set, with the exception of the stack 544, debug 15 address 548 and SRAM address and data register mentioned above. The instruction decoder and operand multiplexer 498 looks at the various registers of set 490 and SRAM 565, decodes the instructions and gathers the operands for operation in the next phase, in particular determining the operands to provide to the ALU 602 below. The outcome of the instruction decoder and operand multiplexer 498 is stored to a number of registers in 20 the second register set 496, including ALU operands 579 and 582, ALU condition code register 580, and a queue channel and command 587 register, which in this embodiment can control thirty-two queues. Several of the registers in set 496 are loaded fairly directly from the instruction register 535 above without substantial decoding by the decoder 498,

including a program control 590, a literal field 589, a test select 584 and a flag select 585.

Other registers such as the file context 522 of the first phase 500 are always stored in a

file context 577 of the second phase 560, but may also be treated as an operand that is

gathered by the multiplexer 572. The stack registers 544 are simply copied in stack

5 register 594. The program counter 540 is incremented 568 in this phase and stored in

register 592. Also incremented 570 is the optional debug address 548, and a load control

575 may be fed from the pipeline 505 at this point in order to allow error control in each

phase, the result stored in debug address 598.

FIG. 22C depicts the third microprocessor phase 600, which includes ALU and

10 queue operations. The ALU 602 includes an adder, priority encoders and other standard

logic functions. Results of the ALU are stored in registers ALU output 618, ALU

condition codes 620 and destination operand results 622. A file context register 616, flag

select register 626 and literal field register 630 are simply copied from the previous phase

560. A test multiplexer 604 is provided to determine whether a conditional jump results

15 in a jump, with the results stored in a test results register 624. The test multiplexer 604

may instead be performed in the first phase 500 along with similar decisions such as fetch

control 511. A stack exchange 608 shifts a stack up or down depending by fetching a

program counter from stack 594 or putting a program counter onto that stack, results of

which are stored in program control 634, program counter 638 and stack 640 registers.

20 The SRAM address may optionally be incremented in this phase 600. Another load

control 610 for another debug address 642 may be forced from the pipeline 505 at this

point in order to allow error control in this phase also. A queue RAM and queue ALU

606 reads from the queue channel and command register 587, stores in SRAM and

rearranges queues, adding or removing data and pointers as needed to manage the queues of data, sending results to the test multiplexer 604 and a queue flags and queue address register 628. Thus the queue RAM and ALU 606 assumes the duties of managing queues for the three processors, a task conventionally performed sequentially by software on a 5 CPU, the queue manager 606 instead providing accelerated and substantially parallel hardware queuing.

The above-described system for protocol processing of data communication results in dramatic reductions in the time required for processing large, connection-based messages. Protocol processing speed is tremendously accelerated by specially designed 10 protocol processing hardware as compared with a general purpose CPU running conventional protocol software, and interrupts to the host CPU are also substantially reduced. These advantages can be provided to an existing host by addition of an intelligent network interface card (INIC), or the protocol processing hardware may be integrated with the CPU. In either case, the protocol processing hardware and CPU 15 intelligently decide which device processes a given message, and can change the allocation of that processing based upon conditions of the message.